WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

PCT

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : H04B — A2

(11) International Publication Number: WO 98/31107

(43) International Publication Date: 16 July 1998 (16.07.98)

(21) International Application Number: PCT/US97/23728

(22) International Filing Date: 30 December 1997 (30.12.97)

(30) Priority Data: 08/779,770 — 7 January 1997 (07.01.97) — US

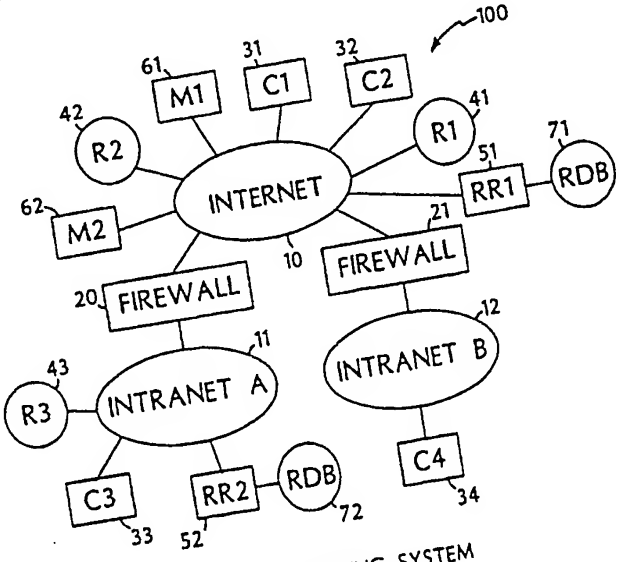
(71)(72) Applicant and Inventor: GIFFORD, David, K. [US/US]; 26 Pigeon Hill Road, Weston, MA 02193 (US).

(74) Agent: WALPERT, Gary, A.; Fish & Richardson, P.C., 225 Franklin Street, Boston, MA 02110–2804 (US).

(81) Designated States: JP, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: REPLICA ROUTING



REPLICA ROUTING SYSTEM

(57) Abstract

The present invention is a new method called replica routing that automatically directs client computers that request a service to a server replica for that service. The replica chosen by replica routing is the replica that is expected to provide the best performance to the client based upon the client's location in the internetwork topology and the estimated performance of the internetwork. In addition, the system and method are designed to permit new replicas to be flexibly added without undue administrative overhead.

REPLICA ROUTING

Background of the Invention

This invention relates in general to an internetwork replica routing system and more particularly relates to a system for directing a client computer to a server replica that is estimated to provide good performance for the client computer.

The recent rapid growth of users of international public packet-switched computer internetworks such as the Internet has created a large demand for the information and services they contain. The replication of services in an internetwork makes it possible for such services to service many users.

Certain known approaches for automatically directing client computers to servers include, for example, round robin DNS and loading balancing DNS, which direct users to one of a number of server replicas in an attempt to balance the load among many servers. In another approach called multiple hostnames, content is spread over multiple servers, each with a separate hostname. Web pages returned to users contain links that point at the replica that has been selected for the user based on load-balancing concerns and replica content considerations. In another approach called Internet load balancing, a hardware component automatically distributes user requests sent to a single IP address to one of a number of server replicas to implement load balancing. Another approach is resonate dispatch that combines load balancing with replica capability to automatically direct users to a replica that is operational, is not overloaded with requests, and contains the requested information.

Summary of the Invention

The invention provides a network server replication system that uses a new method called replica routing to automatically direct a client computer to a server replica

service to contain the replica advertisements for all server replicas. One particular method for replica routing according to the invention allows replica routers to be optionally arranged in a hierarchy, and for replica advertisements to be propagated only part way up the replica router hierarchy. During the replica routing process client requests are automatically sent down the hierarchy until they reach a replica router that is sufficiently knowledgeable about a replica's internetwork location to make an informed replica routing judgment. Thus, not all of the replica advertisements for a particular service have to be contained in a single replica routing server.

A second reason for introducing a hierarchy of replica routers is for security concerns. Since a replica advertisement can contain sensitive information about internetwork characteristics and topology, an organization can choose to create a private replica router inside of a private internetwork (an intranet) to contain this information. In one particular embodiment of the invention, client requests from inside of the intranet will be automatically directed to this private replica router, while client requests from outside of the intranet will use other replica routers that do not need to know the detailed advertisements for replicas they cannot access.

In another aspect of the invention, a client applet can assist in the replica routing process. The client applet can determine certain characteristics of the client internetwork environment, and send these to the replica router as additional information to aid the routing process. The replica router can return more than one replica address to the client applet, and the client applet can then perform empirical performance experiments to choose the best server replica for the use of the client.

replicas 41, 42, 43, 44, such as purchasing of goods or user registration that requires the synchronized updating of shared databases. Replica routers, server replicas, and master servers can be implemented on separate computers as shown, or can share computers. For example, a replica router, a server replica, and a master server can exist on the same computer.

The contents of server replicas 41, 42, 43, 44 can be dynamically maintained by a network-based replication method, such as a master-slave scheme or weighted voting, or replicas can be updated by digital broadcast either over the network or by separate multicast or broadcast channels such as satellite or terrestrial links. Alternatively, replicas can either be partially or totally implemented by media that can be physically distributed such as optical disk.

The software architecture underlying the particular preferred embodiment is based upon the hypertext conventions of the World Wide Web. The Hypertext Markup Language (HTML) document format is used to represent documents and forms, and the Hypertext Transfer Protocol (HTTP) is used between client, replica router, server replica, and master server computers. Documents are named with Uniform Resource Locators (URLs) in the network of computers. A document can be any type of digital data broadly construed, such as multimedia documents that include text, audio, and video, and documents that contain programs. In particular, Java applets and ActiveX controls can be contained in or referenced by documents that allow the capabilities of client computers to be automatically extended by the downloading of new programs.

In addition to documents, server replicas can be used to replicate any type of data, including relational databases, multimedia data, video files, and groupware data. To support access to these datatypes server replicas

subnetworks. The term network number or network identifier is used to refer to the IP address of a network including both its network and subnetwork components.

Fig. 2 shows an example hierarchy 200 of replica routers, with router 201 being a root replica router, and with router 203 being a leaf replica router that contains replica advertisements for server replicas in its network neighborhood. More than one replica router can exist at each level of the hierarchy, and there can be multiple root replica routers. The IP addresses of the root replica routers are bound to the DNS name of the service, such as "www.pathfinder.com."

Before discussing how replica routers operate to direct client computers to server replicas that provide good performance for the client computers, this discussion will describe how a client computer or server replica can "discover" its local internetwork topology. Then this discussion will describe how this technique is used in connection with the replica routing system of the present invention.

Fig. 6 is a flowchart for the discovery of local internetwork topology and performance. All routing protocols that are in use on the internetwork are employed, including the Routing Information Protocol (RIP), External Gateway Protocol (EGP), Boarder Gateway Protocol (BGP), Open Shortest Path First (OPSF), Classless Interdomain Routing (CDIR), and their descendants and follow-ons. At step 905 a client computer (or a server replica) sends a network router solicitation message on all connected networks by broadcast or multicast to learn of nearby network routers (standard network routers are not to be confused with the replica routers according to the invention).

At step 910, network routing table request messages are sent to all of the network routers discovered in step

915, it can be "pinged" to attempt to estimate the network performance from the client to the distant network. At step 925 if a configuration-set maximum number of iterations has not been exceeded, then at step 935 all of the network routers that were named in the routing tables received at step 915 that were not previously explored are assembled, and this set of new routers is used at step 910 to learn more about the network neighborhood. Otherwise, at step 930, internetwork performance discovery is completed, yielding a network performance table that is a list of rows, in which each row contains a network number, a net mask, and an estimate of the performance to that network number.

In an alternative embodiment, the maximum depth (maximum number of iterations) explored for a given network router can depend on the network router (e.g., well-known network routers can have a greater maximum depth). In another alternative embodiment, more than one network performance metric can be utilized (such as bandwidth and latency).

Multiple types of network numbers can be used simultaneously in a network performance table, and thus multiple types of network numbers can be used in replica routing databases, including IP network numbers, IPng (next generation) numbers, and their successors.

In an alternative implementation, the network performance map is extended by using a traceroute utility to perform traceroutes to preconfigured IP addresses and to the IP addresses of server replicas. Traceroute utilities are described in TCP/IP Illustrated, Vol. 1, Chapter 8: "Traceroute Program," Stevens (1994, Addison-Wesley, Reading, Massachusetts). Server replica addresses can be discovered by contacting root replica routers and other replica routers and asking them with a specialized request to transmit their list of server replica and replica router

parent to which it can send an advertisement). If this replica router is not a root replica router, then at step 310 the replica routing database is used to create a new replica summary record that has multiple entries, one for each network number advertised in a replica summary record in the replica routing database. Each entry in the newly created replica summary record includes: a network number, the net mask for that network number, the best performance metric value for that network number that is advertised in a replica summary record by any server replica or replica router, and the timestamp from this best performing entry in the routing database. The newly created replica summary record is marked as being created by a replica router.

At step 315 logic common to the replica routers and server replicas begins, and the new replica summary record can be modified according to operator-specific rules that are specific to the replica router or server replica. Arbitrary alterations to the new replica summary record can be specified, including: the removal of certain networks; the addition of network numbers with specified network masks, performance metric values, and timestamps that can include a "do not expire value"; manual override, by network number, of network masks and performance metric values or replica summary record entries; and removal of replica summary record entries that do not achieve a specified performance metric value. In this way the operator of a server can ensure that the server serves its intended audience, for example by adding intranet network numbers that cannot be seen from outside the intranet's firewall. Next, the replica router or server replica selects a set of parent replica routers. The addresses of the parent replica routers are initialized by looking up the replica routers bound to the service's DNS name (such as "www.pathfinder.com"). Alternatively, the set of parent replica routers can be manually configured for more

After a parent replica router receives the replica advertisement message 335, at step 340 it authenticates the replica advertisement using the public key of the service. Once the advertisement is authenticated, at step 341 a check is made to ensure that the IP address in the advertisement is the same as the source IP address in the header of message 335. If the IP addresses match, control continues at 345, otherwise control continues at 342.

At step 342, if the IP addresses do not match, it means that the replica advertisement has traveled through a firewall (see Fig. 1). The multiple-entry replica summary record in the replica advertisement has a single entry added that includes: the source IP address in the header of message 335, a net mask of all bits "1," a default network metric value, and the current time. This additional entry is added to the summary because the added IP address will be identical to the IP addresses of requests made by clients from behind the same firewall, and thus will match the IP addresses of these client requests. This new replica summary record is marked as having been created by a replica router.

At step 345 the replica advertisement is added to the local routing database at the parent replica router if the advertisement has a more recent timestamp than a previous replica advertisement in the routing database from the same IP address specified in the replica advertisement. Replica advertisements that are superseded by newer advertisements are deleted.

An acknowledgment message is constructed at step 355 that contains the IP address contained in the advertisement, the timestamp, and a digital signature using the service private key. The acknowledgment message 360 is authenticated by the sending server replica or replica router at step 365, and a timer is set at step 370 to refresh the registration information at a configuration-

determined from the entry's replica advertisement). The number N is a configuration parameter. Control then transfers to step 590.

At step 562 the replica router determines the network route and hop-by-hop delay to the client IP address in the header of message 515 using a utility such as traceroute. If the replica router already has the routing and performance information because of a previous execution of step 562 to the same client address it uses this information if the information is not older than a configuration-set maximum time.

At step 565 the IP address of each network router in the network route to the client is looked up in the replica summary records in the replica routing database, starting at the network router closest to the client. Matching is performed using the net mask in each replica summary record entry. If there are no matching entries in the replica routing database, then at step 575 a default set of pre-specified server replicas is made the set of candidate target IP addresses, and all of the default replicas are marked as server replicas. Control is then transferred to step 590.

If there are matching entries in the replica summary records in the routing database then at step 570 each matching replica summary record entry has its advertised network performance added to the network performance estimated from the client to the network router it matched. One way to estimate this performance is to take the round-trip performance observed from the replica router to the client and adjust for the round-trip performance from the replica router to the network router that matched. The N matching replica advertisement entries that contain the best aggregate network performance metric values are selected, and the IP addresses contained in these replica advertisement entries are made the candidate target IP

routing service and can be used for replica routing. For example, certain proposed network routing procedures such as IDPR support network routing servers that can determine the expected network performance of a route between two specified IP addresses in an internetwork. This network service can be directly used in steps 535 to 570 or 575 to pick the server replica in the replica routing database with the best expected performance from the server replica's IP address to the client's IP address.

In an alternative implementation, every root replica router runs on the same computer as a server replica. In this implementation, when no server replica can be found for a particular client's IP address and network location, the replica router directly returns the requested information from its local server replica instead of redirecting the client.

The flowchart in Fig. 5a-5b shows how replica routing can be accomplished with client applets. In step 605 a user activates a link that describes a client applet that mediates access to the target behind the link. In step 610 the applet performs internetwork performance discovery as described in Fig. 6, and at step 615 the applet constructs and sends a replica routing request 620 to one or more root replica routers. The request constructed at step 615 includes the network performance table computed at step 610. Once a client performs step 610, this step does not need to be repeated for a configuration-set interval.

At step 630 a replica router uses the source IP address of message 620, and performs steps 535 to 570 or 575 from Fig. 4a to compute a set of candidate target IP addresses. In the place of step 562, the network performance table computed in step 610 and transmitted in message 620 is used to create an ordered list of network numbers reachable from the client in descending order of

selected for processing in accordance with the original user action at step 605.

In an alternative embodiment, the client applet stores a list of all of the servers and replica routers offered in message 650, and, at step 670, simply constructs a single service request to the server replica or replica router having the best aggregate network performance metric value on the list. In the event that a server or replica router does not respond, the applet will return to the saved list to pick another server address or replica router address to try.

In yet another alternative embodiment, at step 510 or 615 the client constructs a replica routing request that is addressed to a predefined broadcast or multicast address. In this embodiment, replica routers listen for a broadcast or multicast request on this address at 515 or 620. Because multiple replica routers can respond to a broadcast or multicast request, the client can pick the first response, or the response that offers the server replica with the best estimated performance from the client's internetwork location.

Although a system has been described in which replica routers and server replicas all implement a single service (e.g., a single collection of information), generalizations to allow replica routers to function for multiple services can be made by one skilled in the art by introducing appropriate unique service identifiers in messages and database entries and modifying the logic above to include service identifiers.

Novel and improved apparatus and techniques for replica routing have been described herein. It is evident that those skilled in the art may now make numerous uses and modifications of and departures from the specific embodiment described herein without departing from the inventive concept. Consequently, the invention is to be

What is claimed is:

1. An internetwork replica routing system comprising:

a plurality of server replicas, at least one replica router, and at least one client computer interconnected by a communications internetwork;

the client computer being programmed to cause a network request for access to a server replica to be transmitted over the communications internetwork;

at least one replica router being programmed to receive the network request and to calculate a performance metric value for each of at least some of the server replicas that specifies estimated communication performance between the client computer and the server replica, based upon the client computer's location in the internetwork, and being programmed to direct the client computer to at least one server replica that is estimated to provide good performance based upon the client computer's location in the internetwork, the replica router selecting the server replica to which it directs the client computer based on the performance metric values of the server replicas as calculated by the replica router;

the server replica to which the client computer is directed by the replica router being programmed to respond to the network request from the client computer.

2. The system of claim 1 wherein:

the server replicas are programmed to cause server replica advertisements to be sent to the replica router, the advertisements containing information from which the replica router can calculate the performance metric value; and

the replica router is programmed to maintain a database of the server replica advertisements.

8. The system of claim 7 wherein:

at least one the replica routers is programmed to cause a replica router advertisement to be sent to a replica router higher in the hierarchy, the replica router advertisement containing information from which the replica router higher in the hierarchy can calculate the performance metric value; and

the replica router higher in the hierarchy is programmed to store the replica router advertisement in the database of advertisements.

9. The system of claim 8 wherein the replica router higher in the hierarchy is programmed to match the replica router advertisement to its actual source IP address to determine whether the replica router that caused the replica router advertisement to be sent is located behind a firewall.

10. A method of replica routing in a communications internetwork comprising a plurality of server replicas, at least one replica router, and at least one client computer, comprising the steps of:

causing a network request for access to a server replica to be transmitted from the client computer over the communications internetwork;

receiving the network request at at least one replica router;

calculating, at the replica router, a performance metric value for each of at least some of the server replicas that specifies estimated communication performance between the client computer and the server replica, based upon the client computer's location in the internetwork;

directing the client computer to at least one server replica that is estimated to provide good performance based upon the client computer's location in the internetwork,

metric value of at least one network router located in a path from the client computer to the replica router.

15. The method of claim 10, further comprising the step of causing the network request for access to the server replica to be sent from the client computer to the replica router by multicasting or broadcasting the replica routing request over the communications internetwork.

16. The method of claim 10, wherein there are a plurality of replica routers arranged in a hierarchy, and the method further comprises the step of at least one of the replica routers directing the client computer to a server replica that is estimated to provide good performance based upon the client computer's location in the internetwork by directing the client computer to a replica router lower in the hierarchy.

17. The method of claim 16 further comprising the steps of:

causing a replica router advertisement to be sent from one of the replica routers to a replica router higher in the hierarchy, the replica router advertisement containing information from which the replica router higher in the hierarchy can calculate the performance metric value; and

the replica router higher in the hierarchy storing the replica router advertisement in the database of advertisements.

18. The method of claim 17 further comprising the step of matching, at the replica router higher in the hierarchy, the replica router advertisement to its actual source IP address to determine whether the replica router

by the server replica;

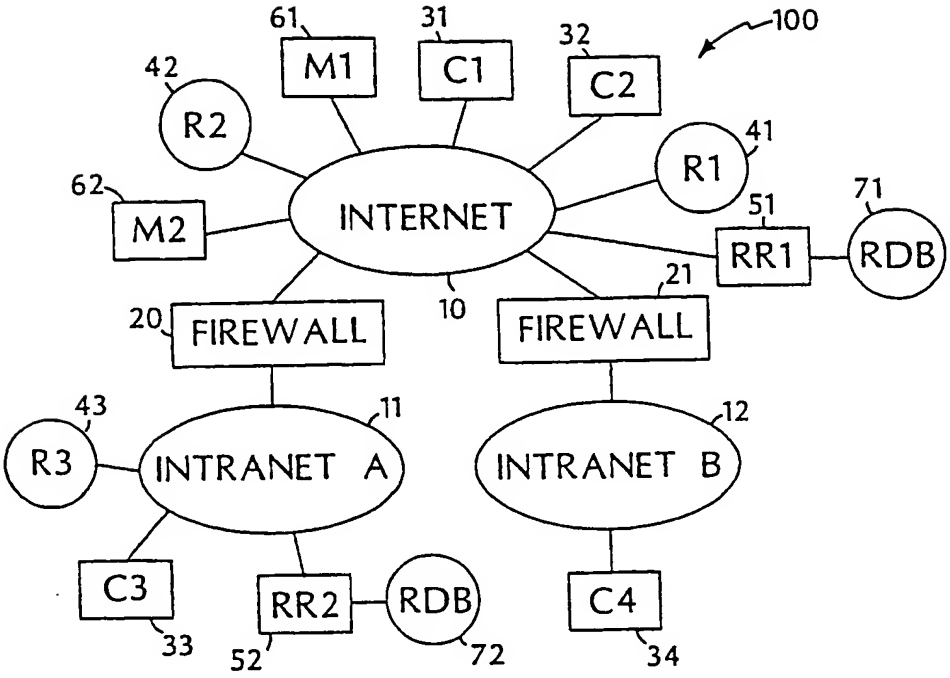
causing a network request for access to a server replica to be transmitted from the client computer over the communications internetwork;

the replica router maintaining a database of the server replica advertisements;

receiving the network request from the client computer at the replica router;

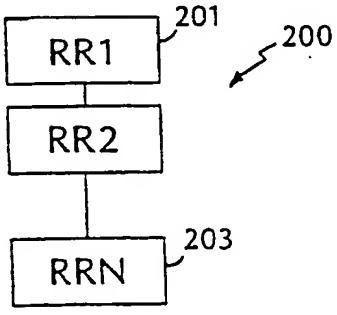
directing the client computer to at least one of the server replicas based upon the relationship between the networks identified in the advertisements in the database and a network in which the client computer is located; and

responding, at the server replica to which the client computer is directed by the replica router, to the network request from the client computer.



REPLICA ROUTING SYSTEM

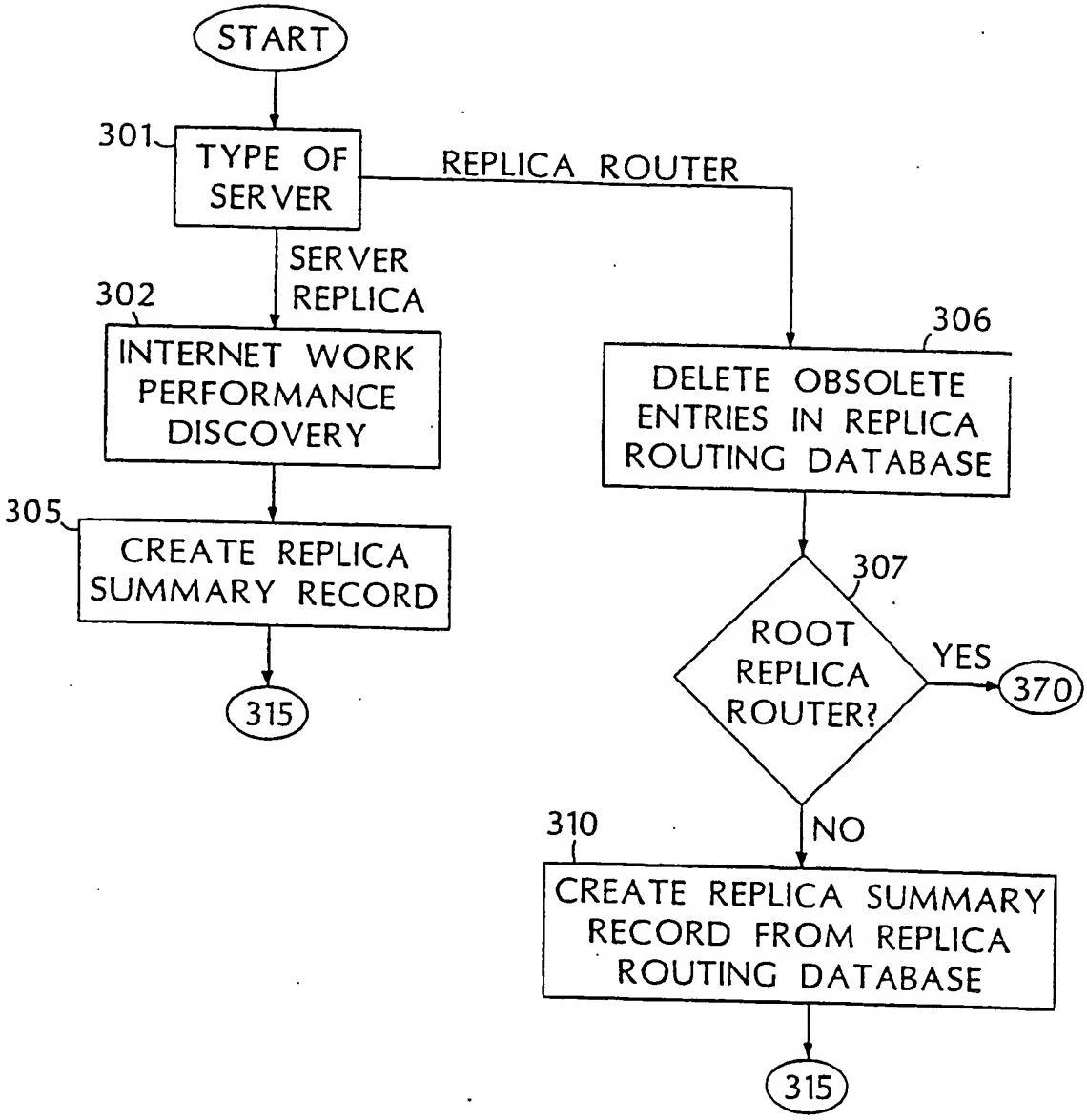
FIG. 1



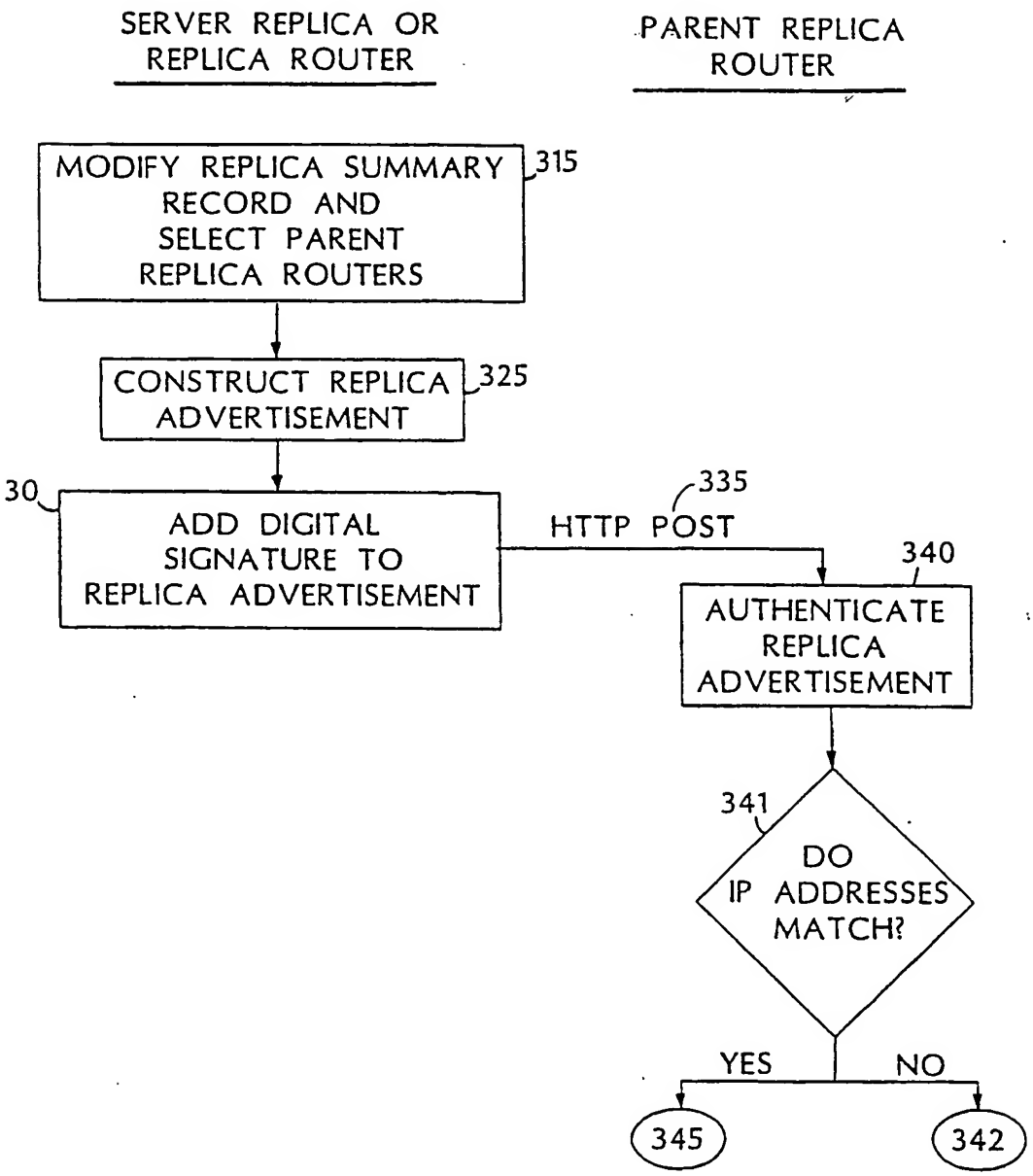
REPLICA ROUTING HIEARCHY

FIG. 2

SERVER REPLICA OR
REPLICA ROUTER

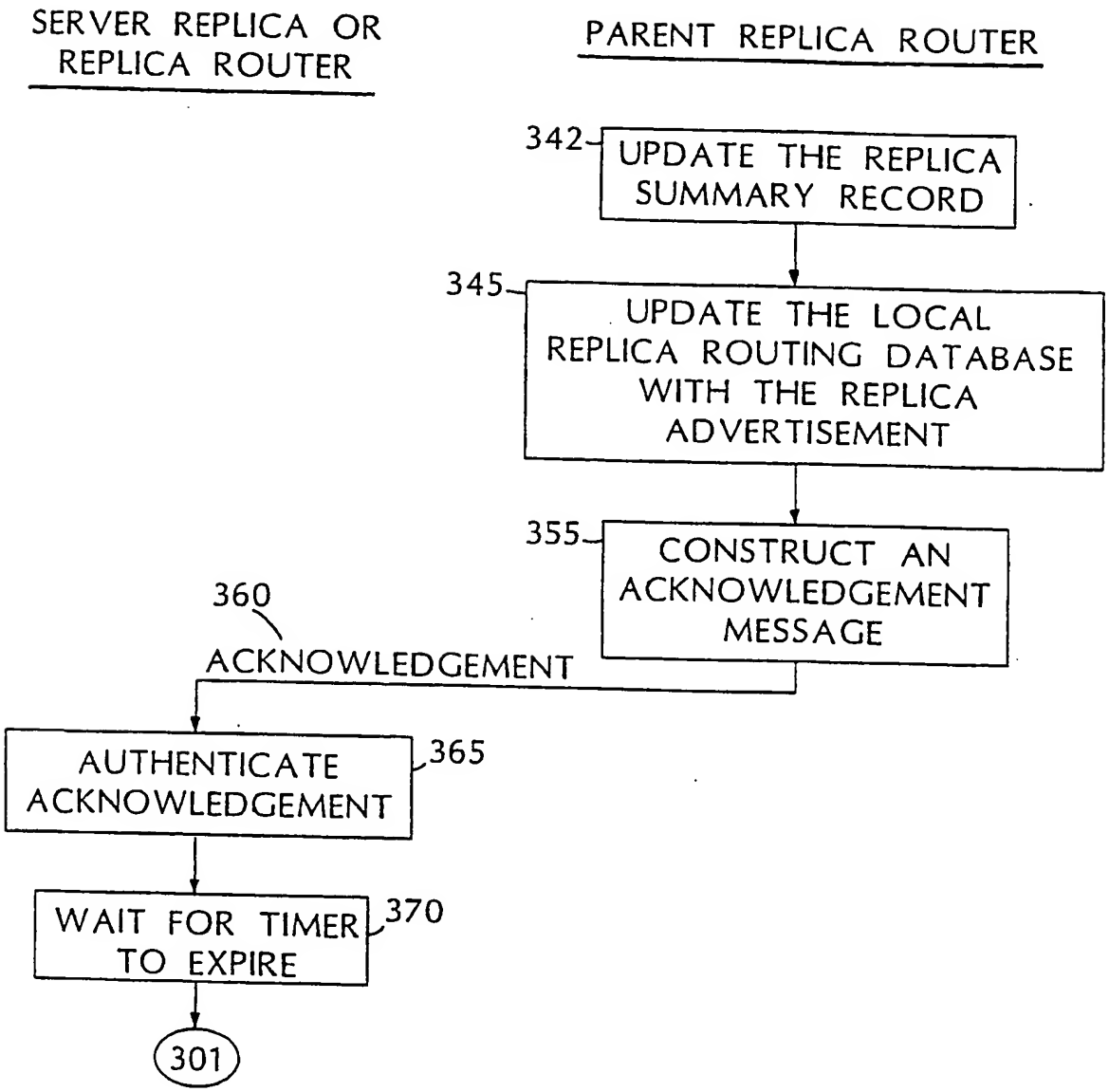


REPLICA ADVERTISEMENT REGISTRATION



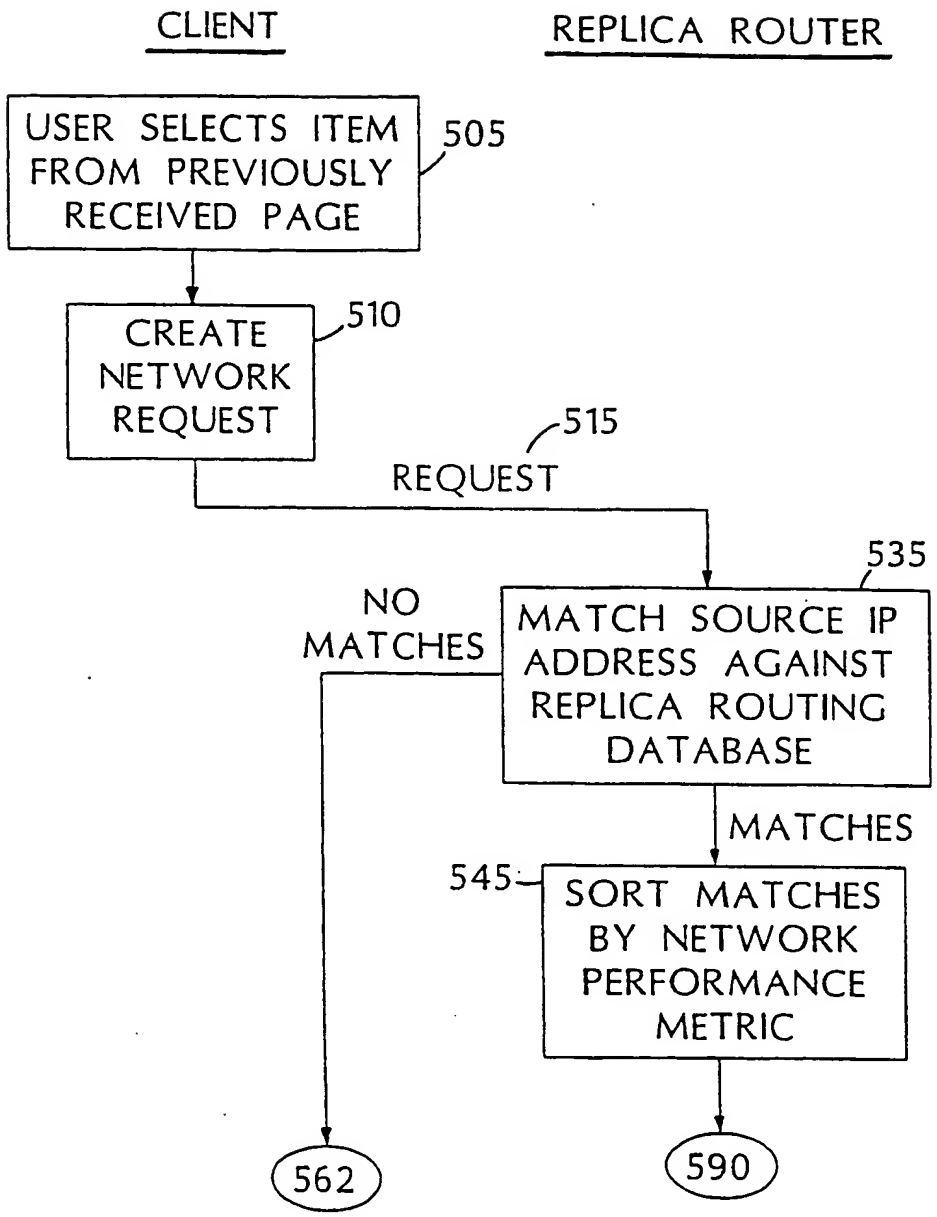
REPLICA ADVERTISEMENT REGISTRATION

FIG. 3B



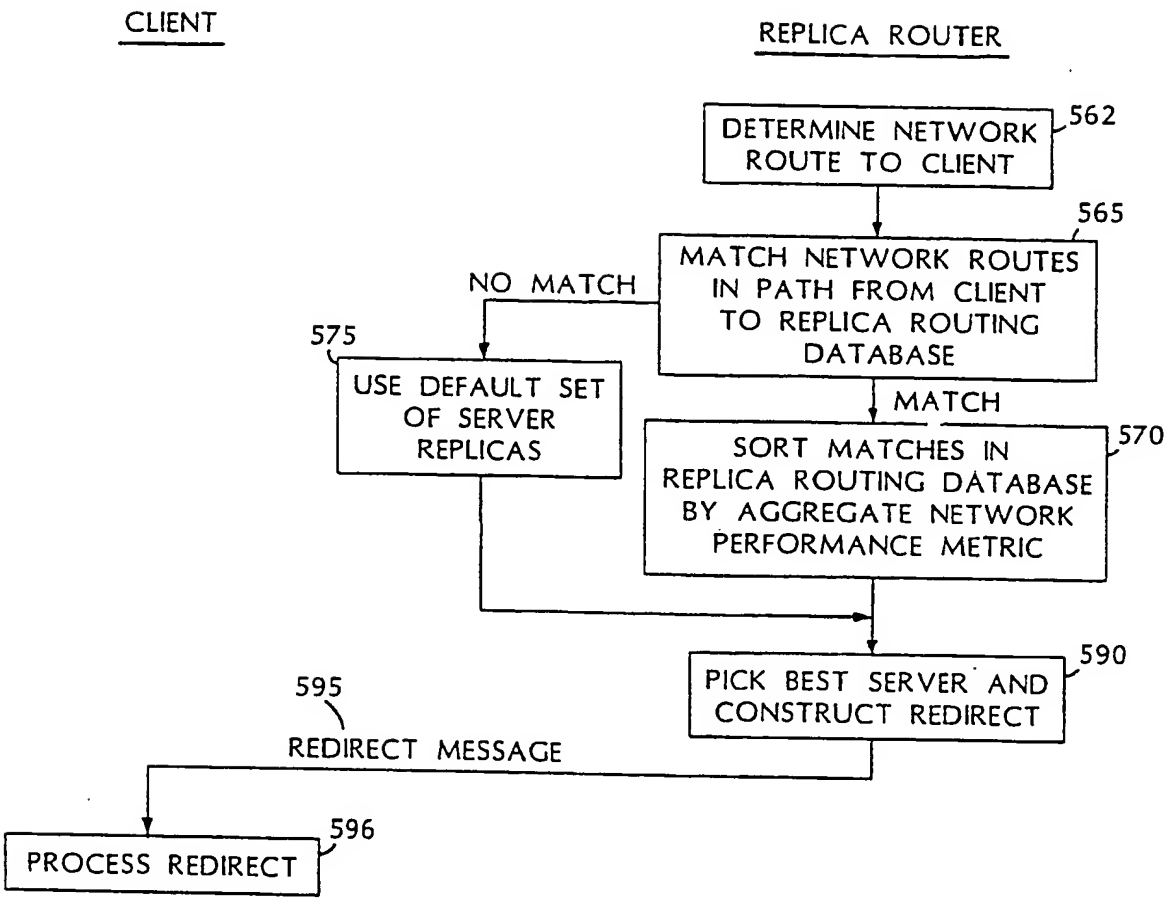
REPLICA ADVERTISEMENT REGISTRATION

FIG. 3C



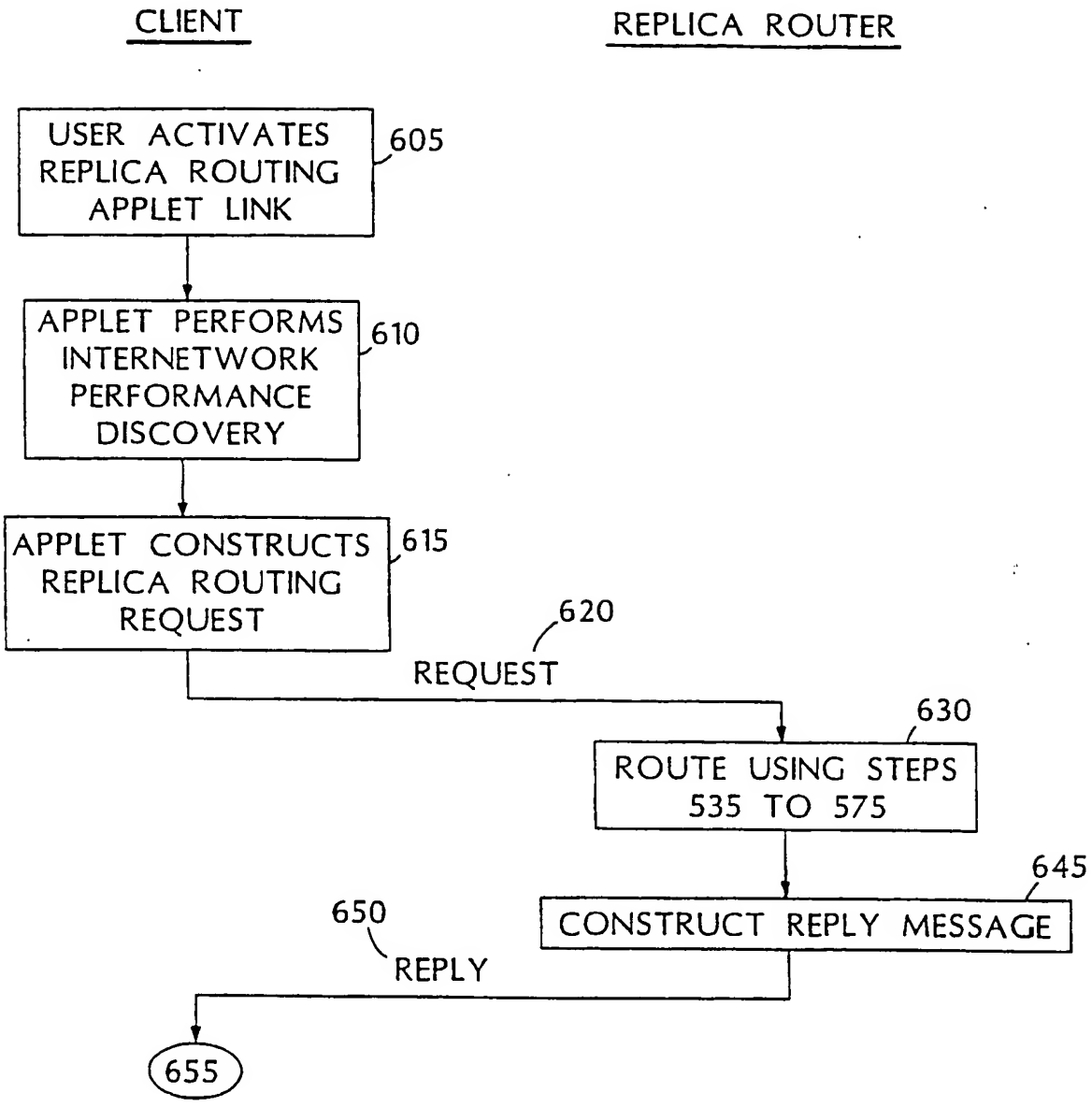
REPLICA ROUTING WITH REDIRECTS

FIG. 4A



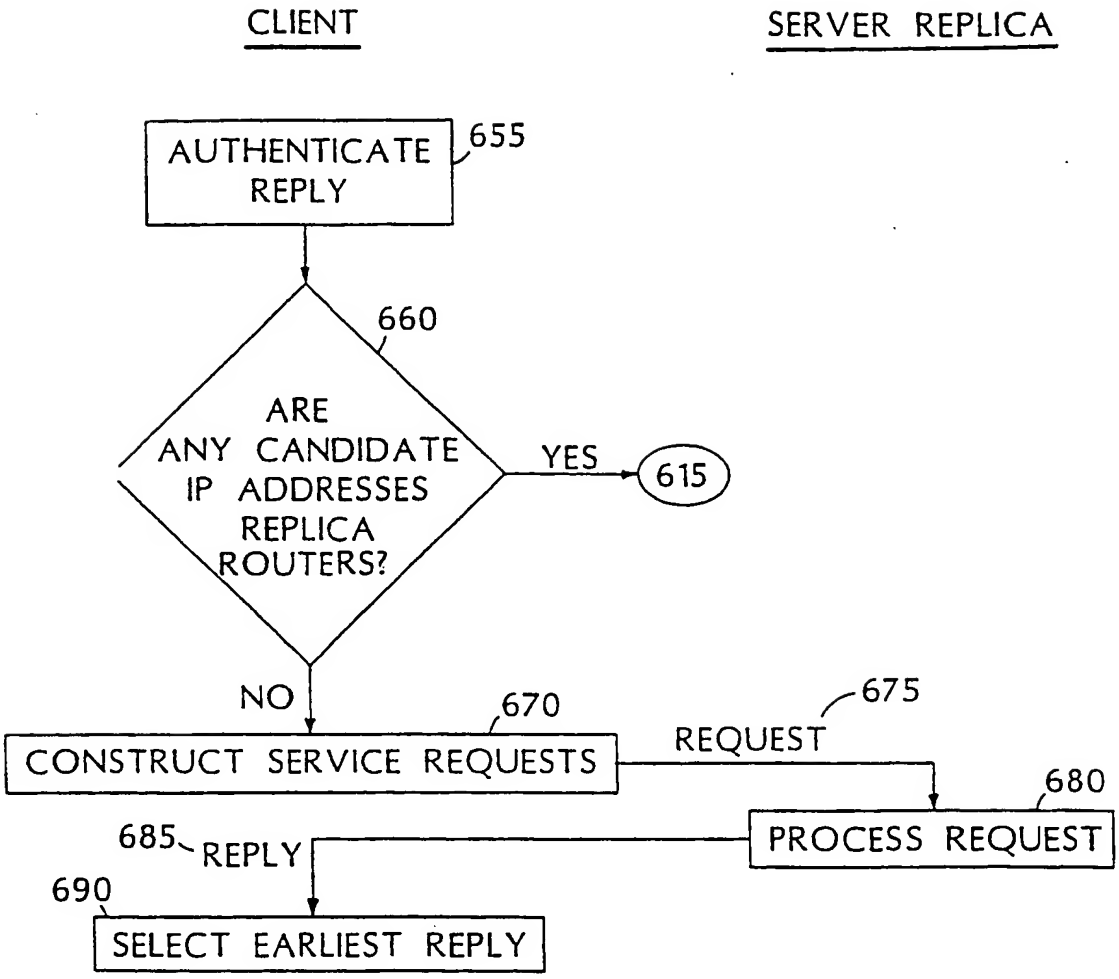
REPLICA ROUTING WITH REDIRECTS

FIG. 4B



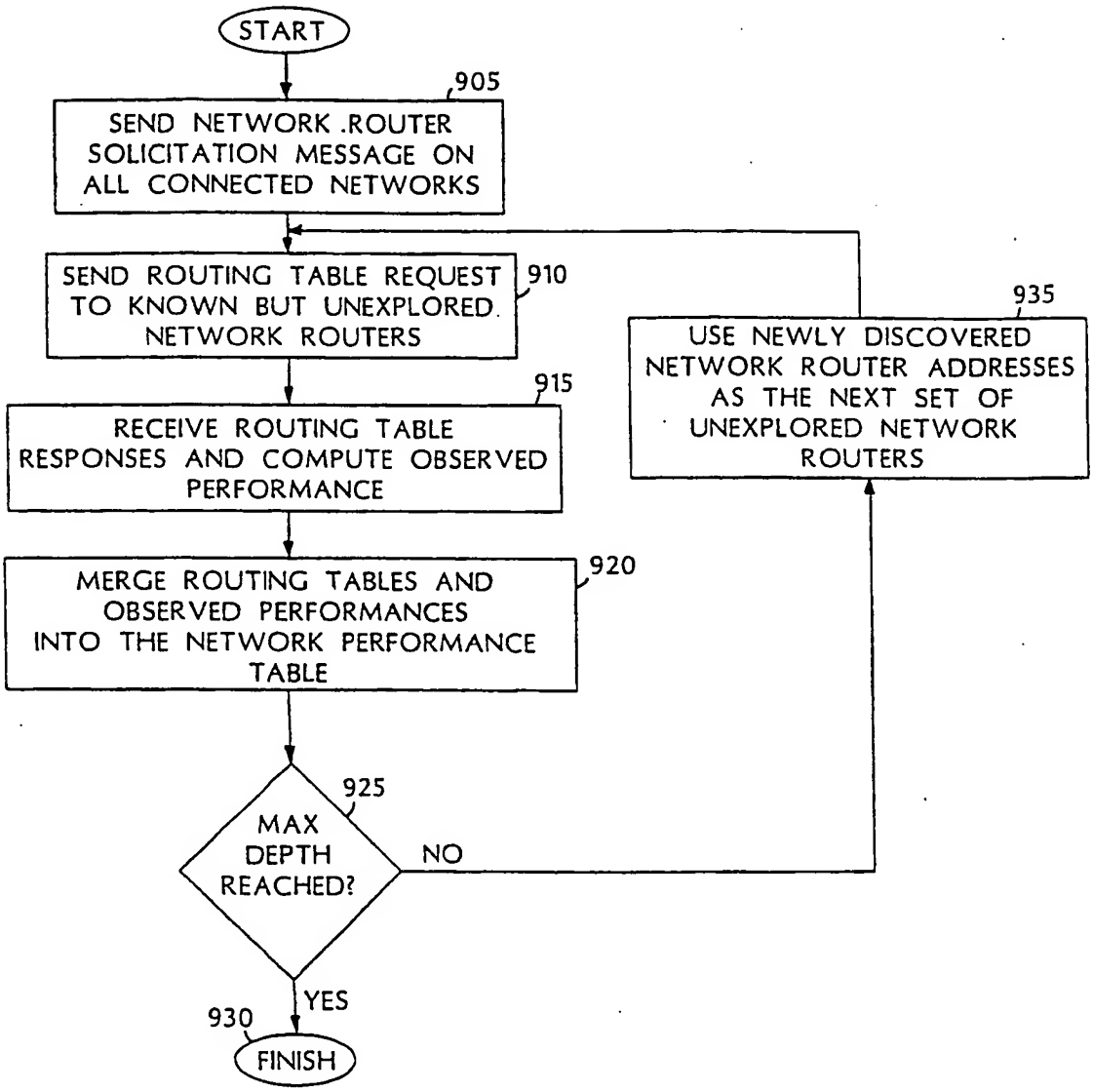
REPLICA ROUTING WITH CLIENT APPLETS

FIG. 5A



REPLICA ROUTING WITH CLIENT APPLETS

FIG. 5B



INTERNETWORK PERFORMANCE DISCOVERY

FIG. 6